MAGVIT: Masked Generative Video Transformer

Year: 2022 | Citations: 323 | Authors: Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang

Abstract

We introduce the MAsked Generative VIdeo Transformer, MAGVIT, to tackle various video synthesis tasks with a single model. We introduce a 3D tokenizer to quantize a video into spatial-temporal visual tokens and propose an embedding method for masked video token modeling to facilitate multi-task learning. We conduct extensive experiments to demonstrate the quality, efficiency, and flexibility of MAGVIT. Our experiments show that (i) MAGVIT performs favorably against state-of-the-art approaches and establishes the best-published FVD on three video generation benchmarks, including the challenging Kinetics-600. (ii) MAGVIT outperforms existing methods in inference time by two orders of magnitude against diffusion models and by 60x against autoregressive models. (iii) A single MAGVIT model supports ten diverse generation tasks and generalizes across videos from different visual domains. The source code and trained models will be released to the public at https://magvit.cs.cmu.edu.